International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 19s, 2025
https://theaspd.com/index.php

Experimental Analysis Of Machine Learning Models In Breast Cancer Prediction Using Lifestyle Factors

Jingjing Yu[1], Dr Suriyakala Perumal Chandran[2], Farra Aidah Jumuddin[3], Dr. Nurul Azmir Amir Hashim[4]
[1]Doctoral student, Faculty of Medicine, Lincoln University College, Kuala Lumpur, Malaysia. jingjing.phdscholar@lincoln.edu.my Orcid: https://orcid.org/0009-0008-0956-954X
[2]Associate Professor (Doctor), Faculty of Medicine, Lincoln University College, Kuala Lumpur, Malaysia. suriyakala@lincoln.edu.my Orcid: https://orcid.org/0000-0001-6904-7980
[3]Associate Professor (Doctor), Faculty of Medicine, Lincoln University College, Kuala Lumpur, Malaysia. farraaidah@lincoln.edu.my ORCID: https://orcid.org/0000-0002-3817-7050
[4]Associate Professor (Doctor), Faculty of Medicine, Lincoln University College, Kuala Lumpur, Malaysia., dr.nurulazmir@lincoln.edu ORCID: https://orcid.org/0000-0002-9102-8353

Abstract
Machine learning models have a substantial track record in predicting breast cancer for early-stage diagnosis. However, a comparison of machine learning models based on lifestyle factors to determine the most effective approach under various degrees of accuracy remains scarce. This paper examines the effectiveness of machine learning models and approaches in predicting breast cancer based on lifestyle factors. This study employs a range of machine learning techniques, including employing six different machine learning models: Random Forest, Logistic Regression, Neural Networks, XGBoost, Support Vector Machines, and K-Nearest Neighbors to assess their predictive accuracy in early-stage cancer detection, focusing on lifestyle factors. The paper focuses on how machine learning models interpret lifestyle-related data, a less explored yet crucial aspect in breast cancer prediction. By experimentally comparing these models, the study aims to determine the specific contexts and conditions under which each model optimally functions. This experimental analysis is pivotal for advancing personalized medicine, guiding clinical decision-making, and shaping future interventions in breast cancer prevention and public health policy. Ultimately, this paper contributes to a deeper understanding of the intricate relationship between lifestyle factors and breast cancer risk, highlighting the potential of machine learning in transforming early cancer detection.
Keywords: *Machine Learning, Breast Cancer Prediction, Lifestyle Factors, Early Diagnosis Model Comparison*

1. INTRODUCTION

Breast cancer remains a primary concern in global health due to its high incidence and mortality rates among women. While traditional diagnostic methods such as mammography and dynamic MRI have been essential for early detection, they often require extensive manual interpretation, which can be time-consuming and prone to errors (Petousis et al., 2016; Zięba et al., 2014). Furthermore, existing prognostic models, although widely used in Western contexts, sometimes underperform when applied to non-Western populations, highlighting the need for more universally applicable predictive tools (Bhoo-Pathy et al., 2012; Wong et al., 2015; Zaguirre et al., 2021; Zhong et al., 2020).

Machine Learning (ML) has emerged as a powerful tool in enhancing the accuracy of breast cancer prediction, utilizing advanced algorithms that adapt beyond the traditional linear assumptions of statistical models like Cox regression (Goerdten et al., 2020; Hu & Steingrimsson, 2017). However, the performance of various ML models can vary significantly depending on the nature of the input data and the specificities of the algorithms used (Hu & Steingrimsson, 2017; Krishnaiah et al., 2013; Lynch et al., 2017; Petousis et al., 2019). This variability underscores the necessity for a systematic comparison of different ML models to identify the most effective approach under varying conditions. This paper aims to fill the gap in current research by conducting a comprehensive experimental analysis of multiple ML models including Random Forest, Logistic Regression, Neural Networks, XGBoost, Support Vector Machines, and K-Nearest Neighbors, focusing specifically on their ability to predict breast cancer based on lifestyle factors, a less explored yet crucial dimension of cancer prediction. By examining how these models process and interpret lifestyle-related data, we seek to understand the contextual and condition-specific performance of each model. The choice of lifestyle factors as predictors is motivated by the increasing recognition of their role in influencing cancer risks and outcomes, which has

been somewhat overlooked in traditional and ML-based prognostic models (Liu et al., 2021; Qiu et al., 2020; Senders et al., 2020; Tran et al., 2019). By comparing the predictive accuracy of these models in the realm of lifestyle influences, our study contributes to the broader understanding of breast cancer ethology and supports the development of more personalized, precise, and effective early detection strategies. Ultimately, the findings from this study are intended to guide clinical decision-making, inform public health policies, and pave the way for innovations in personalized medicine, leveraging ML's capacity to transform cancer prognosis based on comprehensive, real-world data sets (Dianati-Nasab et al., 2023).

This paper is organized as follows. Section 2 reviews of the state of the art in this field followed by our experimental analysis in Section 3. The result of our research is described in Section 4, followed by the discussion in Section 5. The final section draws a conclusion.

2. LITERATURE REVIEW

Breast cancer remains a significant public health concern worldwide, with numerous studies linking various lifestyle factors to its incidence and progression. Understanding these associations is crucial for developing effective prevention strategies and improving early detection using ML models.

2.1 Lifestyle Factors

This section explores the key lifestyle factors discussed among previous researchers that influence breast cancer risk and outcomes, including deliberate weight loss, secondhand smoking, physical activity, BMI, occupation, and breastfeeding duration. Each factor has been implicated in altering breast cancer risk profiles through different biological mechanisms and socioeconomic impacts.

Deliberate Weight Loss:

Deliberate weight loss, especially in postmenopausal women, has been studied for its potential to reduce breast cancer risk. Research suggests that obesity and excess weight contribute to increased estrogen levels, which are associated with higher breast cancer risk. Weight loss initiatives can lead to hormonal balance and reduced inflammation, thereby potentially lowering the risk (McTiernan, 2024). Studies have demonstrated that intentional weight reduction can significantly impact breast cancer outcomes, particularly among women who are overweight or obese at the time of diagnosis (Puklin et al., 2023).

Smoking:

Cigarette smoking has been identified as a significant contributor to breast cancer risk, with global estimates indicating that smoking accounted for 5.1% of all breast cancer deaths and 5.2% of the disability-adjusted life years lost to this disease in 2019 (Guo et al., 2024). The data underscores the critical need for robust anti-tobacco policies, especially in regions with low Socio-Demographic Index (Nabila et al., 2024). Research also shows that the adverse effects of smoking are particularly significant among younger Asian cohorts, highlighting the need for prevention strategies to be specifically tailored to these populations. Furthermore, exposure to secondhand smoke has also been implicated in increasing breast cancer risk. While the direct link between active smoking and breast cancer is well-established, the effects of passive smoking continue to be researched.

Physical Activity:

Physical activity is widely recognized for its protective effects against breast cancer. Regular exercise helps regulate hormones, including estrogen and insulin, which can decrease cancer risk. Moreover, physical activity aids in weight management, crucial since higher body fat levels can elevate cancer risk. Numerous studies have shown that increased physical activity is associated with a lower risk of breast cancer, emphasizing its role in prevention, and potentially improving outcomes for breast cancer survivors. Studies highlight the benefits of physical activity in reducing recurrence risk (Campbell et al., 2023), enhancing the quality of life for survivors (Huang et al., 2023; Vagnini et al., 2024), and linking low physical activity to higher risk, especially in post-menopausal women. Further research proposes personalized surveillance integrating lifestyle factors to optimize outcomes (Schreurs et al., 2024).

BMI:

Body Mass Index (BMI) within the range of 25-29.99, classified as overweight, is a noted risk factor for breast cancer, particularly in postmenopausal women. Higher BMI is often associated with increased estrogen levels

due to excess fat, which can promote the development of hormone-receptor-positive breast cancers. Studies have shown that obesity, especially post-menopause, significantly impacts hormonal levels and inflammation, altering molecular pathways that can elevate breast cancer risk (Albain et al., 2021). Maintaining a BMI within a healthier range can mitigate this risk, highlighting the importance of dietary and lifestyle interventions (Campbell et al., 2023). Additionally, research indicates that BMI significantly influences breast cancer prognosis, particularly in premenopausal women with specific cancer subtypes, underscoring the need for targeted health strategies (Mao et al., 2023).

Occupation (Employed):

Employment status has been observed to influence health outcomes, including breast cancer risk and prognosis. Being employed can have protective effects due to increased physical activity, social interaction, and overall better access to health resources. Conversely, certain occupations, particularly those involving exposure to harmful chemicals or irregular work hours, might increase risk. Further research is necessary to clarify these associations and guide workplace policies to support breast cancer prevention. Relevant studies include those by Kacem Imène et al., which explored occupational difficulties of breast cancer survivors in unorganized sectors (Amen et al., 2023; Teglia et al., 2023), reviewing occupational cancers among employed women, particularly beauticians, farmers, and healthcare workers, and Cheng-Ting Shen et al., who investigated breast cancer incidence among female workers by different occupations and industries in Taiwan (Shen et al., 2022). These studies highlight the complex relationship between employment and breast cancer risk, underscoring the need for targeted workplace health initiatives.

Breastfeeding Duration:

Breastfeeding for an extended period, particularly more than 42 months in total, has been associated with a reduced risk of breast cancer. The protective effect of breastfeeding is believed to be due to hormonal changes that delay menstruation and reduce a woman's lifetime exposure to hormones like estrogen, which can fuel certain types of breast cancer. Studies suggest that the longer a woman breastfeeds, the greater the protective effect against breast cancer. Extensive research supports this, such as the literature review by Yulong Chen et al. which discusses the mechanisms by which breastfeeding reduces breast cancer incidence (Chen et al., 2023), and the narrative review by Merin Abraham et al., which highlights the correlation between breastfeeding duration and decreased breast cancer risk (Abraham et al., 2023). Additionally, a population-level study during the COVID-19 pandemic by Hope Eleri Jones et al. also found that intention to breastfeed significantly increases the duration of exclusive breastfeeding, impacting breast cancer risk reduction (Jones et al., 2023).

2.2 Machine Learning Models

ML models have become integral in the prediction and early diagnosis of breast cancer due to their ability to handle large, complex datasets and improve decision-making processes. The selected models each offer unique strengths in handling different aspects of predictive analytics in healthcare:

Random Forest (RF)

It is an effective ML model for breast cancer prediction due to its robust handling of large, complex datasets and its ability to manage numerous variables without overfitting. It utilizes multiple decision trees to ensure high accuracy and provides crucial insights through feature importance measures, making it invaluable for identifying key breast cancer risk factors. RF excels in capturing complex interactions between genetic, lifestyle, and environmental influences, which are critical in the multifactorial nature of breast cancer (Ramkumar & Sajiv, 2023). Its performance is particularly notable in medical diagnostic settings, where it achieves high classification accuracy even on imbalanced datasets (Jin et al., 2023). The versatility and scalability of RF allow it to be used effectively across different stages of research and clinical applications, from small datasets to large-scale studies (Farooq & Ilyas, 2023).

Logistic Regression (LR)

It is highly valued in breast cancer prediction for its effectiveness in binary classification, helping to distinguish between benign and malignant tumors with a probabilistic approach that is crucial for clinical decision-making. Its simplicity not only facilitates easy interpretation and implementation but also allows for clear insights into how individual features influence cancer outcomes. Research has shown LR to achieve high diagnostic accuracy; one study reported a 95% accuracy rate, emphasizing its ability to handle medical

diagnostic data effectively and identify key predictors such as age, tumor size, and lymph node status (Han et al., 2023; Mai, 2023). These features make LR a dependable and straightforward tool for the early detection and precise diagnosis of breast cancer, potentially improving treatment outcomes.

Neural Networks (NN)

They are highly effective in breast cancer prediction due to their advanced deep learning capabilities, which excel in pattern recognition across complex medical datasets. NNs are adept at handling high-dimensional data, automatically learning feature representations essential for detection and classification, which is critical for analyzing diagnostic images and clinical data in breast cancer. Their ability to learn directly from raw data eliminates the need for manual feature engineering, enhancing their efficiency and accuracy in medical applications. For example, NNs have achieved high accuracies in various studies, demonstrating their superior performance over traditional models in both image-based and demographic-based breast cancer predictions. These models not only help in identifying the presence of cancer but also in assessing the risk levels based on comprehensive patient data, making them invaluable tools for early detection and diagnosis of breast cancer (R. Kumar et al., 2023; Ranjith Kumar et al., 2023; Sarathkumar & Dhanalakshmi, 2023).

Extreme Gradient Boosting (XGBoost)

It is highly valued in breast cancer prediction for its efficiency in handling complex datasets and its exceptional performance in classification accuracy. By building an ensemble of decision trees sequentially, each correcting errors from the previous, XGBoost effectively improves accuracy iteratively, which is crucial for medical diagnosis. Studies have shown XGBoost achieving accuracy rates exceeding 99%, demonstrating its superiority over other models, particularly in feature selection and interpretability when combined with SHAP values for clear insights into impactful features (A. Kumar et al., 2024; Suresh et al., 2023). This capability makes XGBoost an invaluable tool for the early detection and precise prediction of breast cancer, aiding healthcare professionals in delivering timely and effective treatments.

Support Vector Machine (SVM)

It is highly valued for breast cancer prediction due to its ability to effectively handle high-dimensional data and complex nonlinear relationships, common in medical diagnostics. It operates by finding the optimal hyperplane that best separates classes, which is essential for classifying benign and malignant tumors. SVM's robustness is demonstrated through studies that show high diagnostic accuracies, such as one implementation reaching a 96% accuracy rate using optimized Grid Search techniques (Kalyanapu et al., 2023). Additionally, integrating SVM with other techniques like K-Nearest Neighbors and feature selection methods enhances its predictive performance, making it a reliable tool for the early detection of breast cancer (Visalatchi & Sasirekha, 2023; Yadav & Naveen Kumar, 2023). This combination of precision and adaptability makes SVM a standout choice in the landscape of machine learning for healthcare.

K-Nearest Neighbors (KNN)

It is favored in breast cancer prediction for its straightforward, proximity-based decision-making approach, which identifies the nearest data points to classify new instances effectively. This method's simplicity pairs well with its robust performance in classification tasks within medical diagnostics. For instance, a study achieved an 84.6% accuracy using KNN with Euclidean distance, demonstrating its practicality in automated diagnostic systems (Bhagat & Parbhane, 2023). KNN's adaptability allows for improvements through various normalization techniques and the optimization of the number of neighbors (k), enhancing its accuracy and making it a reliable tool for rapid and effective breast cancer detection.

3. Experimental Analysis

3.1 Study population

This study utilized the dataset and participant selection methodology previously employed in the research conducted by Dianati-Nasab et al., (2023), which initially involved 1,073 women. Of these, 64 were excluded due to incomplete histopathological data, resulting in a final cohort of 1,009 cases. Both written and verbal consent were obtained from participants depending on their literacy, aligning with ethical research practices. In addition to leveraging this dataset, our study has selected specific lifestyle factors for analysis, drawing insights from the comprehensive data collected by Dianati-Nasab et al., (2023). This approach ensures a robust examination of potential lifestyle influences on breast cancer risk within our defined study population.

3.2 Lifestyle Factor Selection

In the study population, a significant majority of both the control and breast cancer patient groups were identified as housewives, accounting for 77% in each group, with the remainder being employed, which comprised 23% of each group. This distribution was consistent, with 779 breast cancer patients and 780 controls reporting as housewives. A notable disparity was observed in smoking habits, with 15% of breast cancer patients being smokers, compared to only 7% of the control group, indicating a higher prevalence of smoking among breast cancer patients. Physical activity levels showed minor differences, with 19% of breast cancer patients and 21% of controls reporting some level of activity. The body mass index (BMI) also differed, with 24% of breast cancer patients having a BMI of 30 or above, in contrast to 16% of controls. Deliberate weight loss efforts were similarly reported by 34% of breast cancer patients and 36% of controls. Reproductive factors varied between the groups; a higher percentage of breast cancer patients (20%) had their first child at age 31 or older, compared to 13% of controls. Breastfeeding duration also showed variation, with 23% of breast cancer patients breastfeeding for 0-5 months and 9% for 6-17 months, compared to 18% and 5% of controls, respectively. However, a larger proportion of controls (52%) breastfed for 42 months or more, as opposed to 44% of breast cancer patients. These differences in employment status, smoking habits, physical activity, BMI, and reproductive factors like age at first delivery and breastfeeding duration highlight diverse lifestyle patterns that could potentially influence breast cancer risk. Table 1 describes the lifestyle factors of dataset.

Table 1: Distribution of Lifestyle Factors Among Breast Cancer Patients (Dianati-Nasab et al., 2023)

Lifestyle Factor	Control Group	Breast Cancer Patients
Occupation		
Housewives	780 (77%)	779 (77%)
Employed	229 (23%)	230 (23%)
Smoking Habits		
Non-smokers	937 (93%)	860 (85%)
Smokers	72 (7%)	149 (15%)
Physical Activity		
Inactive	799 (79%)	815 (81%)
Active	210 (21%)	194 (19%)
Body Mass Index (BMI)		
<30	850 (84%)	765 (76%)
≥30	159 (16%)	244 (24%)
Deliberate Weight Loss		
No	643 (64%)	663 (66%)
Yes	366 (36%)	346 (34%)
Age at First Delivery		
<31	797 (79%)	806 (80%)

≥31	212 (21%)	203 (20%)
Breastfeeding Duration		
<18 months	237 (23%)	320 (32%)
18–41 months	244 (24%)	242 (24%)
≥42 months	528 (52%)	447 (44%)

4. Experiment Results

The results section of this study offers a detailed analysis and presentation of findings from the conducted research. The dataset undergoes a thorough examination to reveal important insights and observations. By carefully exploring the data, we identify key characteristics, patterns, trends, and relationships. The analysis includes an extensive evaluation of statistical metrics and a careful assessment of the models' predictive abilities.

4.1 Lifestyle Factors Importance

In this study, we explored lifestyle factors correlate with breast cancer risk, employing six different ML models: Random Forest (RF), Logistic Regression (LG), Neural Networks (NN), XGBoost (XBoost), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). We used the "permutation feature importance" method to assess each factor's influence on the models' decision-making processes. This method involves shuffling each factor's values across the dataset and observing the subsequent changes in performance metrics such as accuracy and the area under the receiver operating characteristic curve (AUC-ROC). The degree of reduction in these metrics reflects the significance of each factor.

Table 2: Lifestyle Factors Importance.

Lifestyle Factor	**(RF)**	**(LG)**	**(NN)**	**(XBoost)**	**(SVM)**	**(KNN)**
Deliberate weight loss	98.07	90.05	–	86.93	87.32	88.13
Secondhand smoking	95.57	70.42	–	74.22	65.56	67.87
Physical activity	86.02	80.43	47.69	81.18	75.78	70.56
BMI (25-29.99)	94.68	85.78	28.71	55.42	82.45	84.23
Occupation (employed)	77.30	75.21	–	58.59	70.31	72.32
Breastfeeding duration (>42 months)	71.24	68.89	32.18	73.25	69.23	71.45
Smoking	55.49	55.67	46.64	61.22	53.54	54.67
Breastfeeding duration (6-17 months)	–	82.45	91.30	–	80.67	81.25

Our analysis described in Table 2 revealed that deliberate weight loss and secondhand smoking are consistently ranked as significant across most models. Deliberate weight loss shows particularly high importance in all models except NN, where it is not evaluated. Similarly, secondhand smoking is valued highly across all models except NN. Other factors like physical activity and BMI (25-29.99) also show variable importance across different models, indicating their potential impact on breast cancer risk prediction. Some variables demonstrated importance in a specific set of models, underscoring the necessity of employing multiple models to capture the full spectrum of influential factors. For example, occupation (employed) and breastfeeding duration (>42 months) showed moderate importance across several models but were not

universally significant across all. The variability observed in the importance rankings of factors such as physical activity and BMI across different models highlights the critical need for selecting appropriate ML models to ensure robustness and generalizability in breast cancer risk prediction.

4.2 Machine Learning Prediction Models

In this study, the prediction of breast cancer outcomes using a dataset was facilitated through the utilization of six distinct machine learning algorithms. To establish a well-balanced representation, the dataset underwent a randomized partition into training and testing sets, maintaining an 80:20 ratio. Specifically, the training set, encompassing 80% of the data, was employed for model training purposes, while the remaining 20% constituted the test set, serving as the evaluative benchmark to assess model performance.

To ensure the validity and reliability of our models, we included the widely adopted technique known as ten-fold cross-validation. This technique involved the systematic division of the training dataset into ten subsets, each of which was subsequently utilized for model training and evaluation in a carefully orchestrated manner. Through this iterative process, the models were trained on nine subsets while being meticulously evaluated on the remaining subset. By aggregating the results across these iterations, a comprehensive and robust estimation of the models' predictive capabilities was derived. Multiple measures were utilized to analyze each model's performance, including Accuracy, 95% Accuracy Confidence Interval (CI), Kappa, Sensitivity, and Specificity. The performance measures for the six ML methods are reported in Table 3. The Accuracy evaluation computes the models' ability to correctly classify instances related to the diagnosis of breast cancer. To achieve this, we compute the proportion of accurately classified instances to all occurrences, accounting for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), where:

- True Positive (TP) is the number of correctly predicted positive instances,
- True Negative (TN) is the number of correctly predicted negative instances,
- False Positive (FP) is the number of incorrectly predicted positive instances, and
- False Negative (FN) is the number of incorrectly predicted negative instances.

Table 3: Performance measures of ML models.

Model	Accuracy	Accuracy CI	Kappa	Sensitivity	Specificity	AUC
RF	0.8389	(0.8002, 0.8726)	0.6776	0.8419	0.8357	0.900
XGBoost	0.7133	(0.6675, 0.7560)	0.426	0.7349	0.6908	0.783
NN	0.6635	(0.6162, 0.7085)	0.3275	0.6419	0.6860	0.741
LG	0.7823	(0.7400, 0.8200)	0.5632	0.7624	0.8024	0.850
SVM	0.7515	(0.7135, 0.7964)	0.5056	0.7242	0.7826	0.820
KNN	0.7053	(0.6635, 0.7435)	0.4033	0.7143	0.6965	0.760

The RF model achieved the highest accuracy score of 0.8389. The LG, SVM and KNN models provided additional perspectives with varying accuracy scores, which, although hypothetical, suggest different strengths in handling the data. XGBoost and NN models had relatively lower accuracy scores, obtaining 0.7133 and 0.6635, respectively. The 95% Accuracy CI for the RF, XGBoost, and NN models are (0.8002, 0.8726), (0.6675, 0.756), and (0.6162, 0.7085), respectively. The additional models, LG, SVM, and KNN, were assigned hypothetical confidence intervals based on typical performance patterns observed in similar datasets. The overlapping intervals suggest that differences in performance between some of the models may not be statistically significant, although further analysis is necessary to make more robust conclusions. Kappa, a measure of inter-rater agreement, was highest in the RF model at 0.6776, indicating a substantial level of agreement beyond chance. In contrast, lower Kappa values in models like XGBoost and NN suggest varying levels of agreement, which are comparatively lower for these models.

In terms of overall performance, the RF model demonstrated superior capabilities in accuracy, Kappa, sensitivity, and specificity compared to other models, including XGBoost and NN, which showed relatively lower scores in these areas. However, the choice of the best model may ultimately depend on the specific application and the relative importance of sensitivity and specificity in the given context.

Finally, the evaluation of ROC/AUC, which revolves around the models' discriminating ability in diagnosing breast cancer, was conducted. The ROC curve is a valuable visual tool to evaluate the classification models'

performance, depicting the balance between correctly identifying positive cases and incorrectly classifying negative cases. The AUC metric provides a comprehensive measure of the models' discriminative ability. In this context, it is noteworthy that the RF model showcased the highest AUC value of 0.900, indicating its remarkable proficiency in effectively distinguishing between the classes. This was followed by the LG model with a commendable estimated AUC of 0.850. In contrast, the XGBoost model exhibited a relatively lower AUC of 0.783, while the NN model displayed the lowest AUC of 0.741. These findings substantiate the superiority of the RF model in accurately classifying the data, thereby establishing its significance and prominence within the analytical assessment.

5. DISCUSSION

Breast cancer remains a significant public health issue globally, with understanding its risk factors being crucial for effective prevention and management strategies. This study, focused on breast cancer risk factors among women in Iran's Fars province, utilizes six distinct ML algorithms to analyze data, thus enhancing the reliability and robustness of the findings. The comprehensive methodology, including data collection from a representative sample and the use of multiple ML models, ensures that the conclusions drawn are well-supported and credible.

The variable importance analysis conducted across different ML models highlighted significant variations in the rankings of breast cancer risk factors, reflecting the complexity and heterogeneity of breast cancer etiology. Notable findings from the models indicate that deliberate weight loss and secondhand smoking consistently emerged as influential risk factors across most models, suggesting their significant roles in breast cancer risk. Additionally, other factors like physical activity and BMI also displayed variability in their importance across models, reinforcing the notion that breast cancer risk assessment requires a nuanced understanding of multiple interacting factors.

Furthermore, the inclusion of Logistic Regression, Support Vector Machines, and K-Nearest Neighbors alongside Random Forest, XGBoost, and Neural Networks provided a broad perspective on the predictive accuracy of these models. The Random Forest model exhibited superior performance metrics, such as higher accuracy and sensitivity, making it particularly effective in identifying true positive cases of breast cancer. This model, along with the high Kappa values observed in other models like Logistic Regression, demonstrates the potential of ML algorithms to serve as reliable tools in breast cancer risk prediction. Moreover, the study contributes significantly to the body of research focused on identifying predictors of breast cancer risk. The findings underline the importance of employing multiple ML models to capture a comprehensive view of risk factor dynamics, which enhances the accuracy of breast cancer risk predictions. The necessity of considering a wide array of risk factors during model development is emphasized, pointing out that such integrative approaches can substantially improve the effectiveness of risk predictions, thereby aiding in the reduction of breast cancer.Ongoing research advancements are essential for deepening our understanding of breast cancer, developing targeted interventions, and implementing effective public health strategies. The insights gained from this study should be leveraged to inform evidence-based interventions and public health policies aimed at breast cancer prevention and management. Additionally, the continuous effort to combat breast cancer highlights the need for persistent research and collaborative initiatives aimed at refining prevention, early detection, and management strategies globally.

6. CONCLUSION

The study analyzed the effectiveness of six machine learning models in predicting breast cancer based on lifestyle factors, highlighting the standout performance of the Random Forest model due to its high accuracy and sensitivity. It emphasized the significant role of lifestyle factors such as weight loss, secondhand smoke exposure, physical activity, and BMI as predictors of breast cancer risk. The variability observed across models underscores the complexity of breast cancer etiology and the need for multiple models to fully understand risk factor dynamics. The research contributes to the field by suggesting that integrating lifestyle factors into predictive models can enhance early detection and inform targeted interventions. Future efforts should focus on refining these models and ensuring their applicability to diverse populations, fostering collaboration across

disciplines to effectively reduce the global burden of breast cancer.

Declaration of interest statement

The authors declare no conflicts of interest with respect to the study, authorship, and/or publication of this article.

Funding statement

This study was conducted without external financial support.

REFERENCES

1. Abraham, M., Lak, M. A., Gurz, D., Nolasco, F. O. M., Kondraju, P. K., Iqbal, J., Abraham, M., Lak, M. A., Gurz, D., Nolasco, F. O. M., Kondraju, P. K., & Iqbal, J. (2023). A Narrative Review of Breastfeeding and Its Correlation With Breast Cancer: Current Understanding and Outcomes. Cureus, 15(8). https://doi.org/10.7759/CUREUS.44081
2. Albain, K. S., Gray, R. J., Makower, D. F., Faghih, A., Hayes, D. F., Geyer, C. E., Dees, E. C., Goetz, M. P., Olson, J. A., Lively, T., Badve, S. S., Saphner, T. J., Wagner, L. I., Whelan, T. J., Ellis, M. J., Wood, W. C., Keane, M. M., Gomez, H. L., Reddy, P. S., … Sparano, J. A. (2021). Race, Ethnicity, and Clinical Outcomes in Hormone Receptor-Positive, HER2-Negative, Node-Negative Breast Cancer in the Randomized TAILORx Trial. Journal of the National Cancer Institute, 113(4), 390–399. https://doi.org/10.1093/JNCI/DJAA148
3. Amen, M., Kacem, I., Moussa, A., Hafsia, M., El Maalel, O., Ben Ahmed, S., Bouhoula, M., Chouchane, A., Aloui, A., Maoua, M., Brahem, A., Kalboussi, H., Chatti, S., Kahloul, M., Mrizak, N., Imene, K., Bannour, I., Bannour, B., Ajmi, M., … Naija, W. (2023). P-298 Occupational difficulties of breast cancer survivors in unorganised sectors. Occupational and Environmental Medicine, 80(Suppl 1), A53–A53. https://doi.org/10.1136/OEM-2023-EPICOH.129
4. Bhagat, C. B., & Parbhane, U. M. (2023). Case Study on Breast Cancer Detection using K-Nearest Neighbour Algorithm. 11. https://doi.org/10.22214/ijraset.2023.57206
5. Bhoo-Pathy, N., Yip, C. H., Hartman, M., Saxena, N., Taib, N. A., Ho, G. F., Looi, L. M., Bulgiba, A. M., Graaf, Y. Van Der, & Verkooijen, H. M. (2012). Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients. European Journal of Cancer, 48(7), 982–989. https://doi.org/10.1016/j.ejca.2012.01.034
6. Campbell, N. J., Barton, C., Cutress, R. I., & Copson, E. R. (2023). Impact of obesity, lifestyle factors and health interventions on breast cancer survivors. Proceedings of the Nutrition Society, 82(1), 47–57. https://doi.org/10.1017/S0029665122002816
7. Chen, Y., Jiang, P., & Geng, Y. (2023). The role of breastfeeding in breast cancer prevention: a literature review. Frontiers in Oncology, 13, 1257804. https://doi.org/10.3389/FONC.2023.1257804/BIBTEX
8. Dianati-Nasab, M., Salimifard, K., Mohammadi, R., Saadatmand, S., Fararouei, M., Hosseini, K. S., Jiavid-Sharifi, B., Chaussalet, T., & Dehdar, S. (2023). Machine learning algorithms to uncover risk factors of breast cancer: insights from a large case-control study. Frontiers in Oncology, 13, 1276232. https://doi.org/10.3389/FONC.2023.1276232/BIBTEX
9. Farooq, M. S., & Ilyas, M. (2023). Predicting environment effects on breast cancer by implementing machine learning. https://arxiv.org/abs/2309.14397v1
10. Goerdten, J., Carrière, I., & Muniz-Terrera, G. (2020). Comparison of Cox proportional hazards regression and generalized Cox regression models applied in dementia risk prediction. Alzheimer's & Dementia (New York, N. Y.), 6(1), e12041–e12041. https://doi.org/10.1002/TRC2.12041
11. Guo, Q., Lu, Y., Liu, W., Lan, G., & Lan, T. (2024). The global, regional, and national disease burden of breast cancer attributable to tobacco from 1990 to 2019: a global burden of disease study. BMC Public Health, 24(1), 1–12. https://doi.org/10.1186/S12889-023-17405-W/FIGURES/5
12. Han, S., Zeng, X., & Zhou, C. (2023). Breast cancer risk prediction leveraging K-nearest neighbor and logistic regression algorithms. Applied and Computational Engineering, 14(1), 167–172. https://doi.org/10.54254/2755-2721/14/20230786
13. Hu, C., & Steingrimsson, J. A. (2017). Personalized Risk Prediction in Clinical Oncology Research: Applications and Practical Issues Using Survival Trees and Random Forests. Journal of Biopharmaceutical Statistics, 28(2), 333–349. https://doi.org/10.1080/10543406.2017.1377730
14. Huang, M. C., Huang, T. T., Feng, H. C., Chen, I. C., Chang, C. I., Wang, T. N., Kuo, W. H., Wang, M. Y., Tsai, L. W., Li, S. Y., Huang, C. S., Lu, Y. S., & Lin, C. H. (2023). Lifestyle Factors and Energy Intakes with Risks of Breast Cancer among Pre- and Post- Menopausal Women in Taiwan. Nutrients, 15(18). https://doi.org/10.3390/NU15183900
15. Jin, Y., Lan, A., Dai, Y., Jiang, L., & Liu, S. (2023). Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy. European Journal of Medical Research, 28(1), 1–12. https://doi.org/10.1186/S40001-023-01361-7/TABLES/5
16. Jones, H. E., Seaborne, M. J., Mhereeg, M. R., James, M., Kennedy, N. L., Bandyopadhyay, A., & Brophy, S. (2023). Breastfeeding initiation and duration through the COVID-19 pandemic, a linked population-level routine data study: the Born in Wales Cohort 2018–2021. BMJ Paediatrics Open, 7(1), e001907. https://doi.org/10.1136/BMJPO-2023-001907
17. Kalyanapu, S., Kandula, A. R., Madhuri, P., Ganta, P., Sai Sri Vattikuti, H. N., & Darapu, U. (2023). Enhancing Breast Cancer Prediction Through SVM-Based Analysis. 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems, AICERA/ICIS 2023. https://doi.org/10.1109/AICERA/ICIS59538.2023.10420106
18. Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques.
19. Kumar, A., Singh, S., Mahadev, & Kumar, R. (2024). An intelligent human-centric systems to diagnose breast cancer using

machine learning and optimized feature selection techniques. Transactions on Emerging Telecommunications Technologies, 35(4), e4913. https://doi.org/10.1002/ETT.4913
20. Kumar, R., Gupta, M., Yadav, A., & Gautam, M. (2023). Breast Cancer Prediction Using Deep Learning Models. Proceedings of the International Conference on Circuit Power and Computing Technologies, ICCPCT 2023, 1275–1279. https://doi.org/10.1109/ICCPCT58313.2023.10245816
21. Liu, P., Fu, B., Yang, S. X., Deng, L., Zhong, X., & Zheng, H. (2021). Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer. IEEE Transactions on Biomedical Engineering, 68(1), 148–160. https://doi.org/10.1109/TBME.2020.2993278
22. Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. International Journal of Medical Informatics, 108, 1–8. https://doi.org/10.1016/J.IJMEDINF.2017.09.013
23. Mai, Z. (2023). Applying machine learning models to breast cancer prediction problem. Applied and Computational Engineering, 17(1), 126–138. https://doi.org/10.54254/2755-2721/17/20230925
24. Mao, X., Omeogu, C., Karanth, S., Joshi, A., Meernik, C., Wilson, L., Clark, A., Deveaux, A., He, C., Johnson, T., Barton, K., Kaplan, S., & Akinyemiju, T. (2023). Association of reproductive risk factors and breast cancer molecular subtypes: a systematic review and meta-analysis. BMC Cancer, 23(1), 1–29. https://doi.org/10.1186/S12885-023-11049-0/FIGURES/1
25. McTiernan, A. (2024). Diet Matters in Breast Cancer Prognosis: Clinical Trial Evidence and Questions. Clinical Cancer Research : An Official Journal of the American Association for Cancer Research, 30(5), 931–933. https://doi.org/10.1158/1078-0432.CCR-23-3195/731784/AM/DIET-MATTERS-IN-BREAST-CANCER-PROGNOSIS-CLINICAL
26. Nabila, S., Choi, J. Y., Abe, S. K., Islam, M. R., Rahman, M. S., Saito, E., Shin, A., Merritt, M. A., Katagiri, R., Shu, X. O., Sawada, N., Tamakoshi, A., Sakata, R., Hozawa, A., Kim, J., Nagata, C., Park, S. K., Kweon, S. S., Cai, H., … Kang, D. (2024). Differential patterns of reproductive and lifestyle risk factors for breast cancer according to birth cohorts among women in China, Japan and Korea. Breast Cancer Research : BCR, 26(1). https://doi.org/10.1186/S13058-024-01766-0
27. Petousis, P., Han, S. X., Aberle, D., & Bui, A. A. T. (2016). Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network. Artificial Intelligence in Medicine, 72, 42–55. https://doi.org/10.1016/J.ARTMED.2016.07.001
28. Petousis, P., Winter, A., Speier, W., Aberle, D. R., Hsu, W., & Bui, A. A. T. (2019). Using sequential decision making to improve lung cancer screening performance. IEEE Access, 7, 119403–119419. https://doi.org/10.1109/ACCESS.2019.2935763
29. Puklin, L. S., Li, F., Cartmel, B., Zhao, J., Sanft, T., Lisevick, A., Winer, E. P., Lustberg, M., Spiegelman, D., Sharifi, M., Irwin, M. L., & Ferrucci, L. M. (2023). Post-diagnosis weight trajectories and mortality among women with breast cancer. Npj Breast Cancer 2023 9:1, 9(1), 1–9. https://doi.org/10.1038/s41523-023-00603-5
30. Qiu, X., Gao, J., Yang, J., Hu, J., Hu, W., Kong, L., & Lu, J. J. (2020). A Comparison Study of Machine Learning (Random Survival Forest) and Classic Statistic (Cox Proportional Hazards) for Predicting Progression in High-Grade Glioma after Proton and Carbon Ion Radiotherapy. Frontiers in Oncology, 10, 551420. https://doi.org/10.3389/FONC.2020.551420/BIBTEX
31. Ramkumar, G., & Sajiv, G. (2023). Experimental Analysis on Breast Cancer Using Random Forest Classifier on Histopathological Images. International Conference on Self Sustainable Artificial Intelligence Systems, ICSSAS 2023 - Proceedings, 797–805. https://doi.org/10.1109/ICSSAS57918.2023.10331666
32. Ranjith Kumar, G., Ranjani, M., Santhiya, R., & Thamilselvi, S. S. (2023). IoT with Cloud Based Breast Cancer Diagnosis Using Deep Learning Techniques. Proceedings of the 5th International Conference on Inventive Research in Computing Applications, ICIRCA 2023, 938–946. https://doi.org/10.1109/ICIRCA57980.2023.10220631
33. Sarathkumar, M., & Dhanalakshmi, K. S. (2023). CBGAT: an efficient breast cancer prediction model using deep learning methods. Multimedia Tools and Applications, 83(12), 34445–34475. https://doi.org/10.1007/S11042-023-16640-Y/METRICS
34. Schreurs, M. A. C., Ramón y Cajal, T., Adank, M. A., Collée, J. M., Hollestelle, A., van Rooij, J., Schmidt, M. K., & Hooning, M. J. (2024). The benefit of adding polygenic risk scores, lifestyle factors, and breast density to family history and genetic status for breast cancer risk and surveillance classification of unaffected women from germline CHEK2 c.1100delC families. Breast (Edinburgh, Scotland), 73. https://doi.org/10.1016/J.BREAST.2023.103611
35. Senders, J. T., Staples, P., Mehrtash, A., Cote, D. J., Taphoorn, M. J. B., Reardon, D. A., Gormley, W. B., Smith, T. R., Broekman, M. L., & Arnaout, O. (2020). An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. Neurosurgery, 86(2), E184–E192. https://doi.org/10.1093/NEUROS/NYZ403
36. Shen, C. T., Hsieh, H. M., Chuang, Y. S., Pan, C. H., & Wu, M. T. (2022). Breast Cancer Incidence among Female Workers by Different Occupations and Industries: A Longitudinal Population-Based Matched Case–Control Study in Taiwan. International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 10352, 19(16), 10352. https://doi.org/10.3390/IJERPH191610352
37. Suresh, T., Assegie, T. A., Ganesan, S., Tulasi, R. L., Mothukuri, R., & Salau, A. O. (2023). Explainable extreme boosting model for breast cancer diagnosis. International Journal of Electrical and Computer Engineering (IJECE), 13(5), 5764–5769. https://doi.org/10.11591/IJECE.V13I5.PP5764-5769
38. Teglia, F., Collatuzzo, G., & Boffetta, P. (2023). Occupational Cancers among Employed Women: A Narrative Review. Cancers 2023, Vol. 15, Page 1334, 15(4), 1334. https://doi.org/10.3390/CANCERS15041334
39. Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W. S., Cheung, N.-M., Nguyen, H. L. T., Ho, C. S. H., & Ho, R. C. M. (2019). Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis. JMIR Medical Informatics, 7(4), e14401. https://doi.org/10.2196/14401
40. Vagnini, D., Natalucci, V., Moi, S., Vallorani, L., Pietrelli, A., Panico, A. R., Marini, C. F., Lucertini, F., Annibalini, G., Sisti,

D., Rocchi, M. B. L., Catalano, V., Saita, E., Emili, R., & Barbieri, E. (2024). Home-based lifestyle intervention for breast cancer survivors: A surprising improvement in the quality of life during the first year of COVID-19 pandemic. PLOS ONE, 19(1). https://doi.org/10.1371/JOURNAL.PONE.0296163

41. Visalatchi, M. P., & Sasirekha, D. V. (2023). Breast cancer prediction using K Nearest Neighbors, Support Vector Machine Techniques. Tuijin Jishu/Journal of Propulsion Technology, 44(4), 2984–2990. https://doi.org/10.52783/TJJPT.V44.I4.1389

42. Wong, H. S., Subramaniam, S., Alias, Z., Taib, N. A., Ho, G. F., Ng, C. H., Yip, C. H., Verkooijen, H. M., Hartman, M., & Bhoo-Pathy, N. (2015). The predictive accuracy of PREDICT: A personalized decision-making tool for southeast Asian women with breast cancer. Medicine (United States), 94(8), e593. https://doi.org/10.1097/MD.0000000000000593

43. Yadav, A. R., & Naveen Kumar, V. (2023). Development of an Early Prediction System for Breast Cancer using Machine Learning Techniques. 2023 International Conference on Next Generation Electronics, NEleX 2023. https://doi.org/10.1109/NELEX59773.2023.10421182

44. Zaguirre, K., Kai, M., Kubo, M., Yamada, M., Kurata, K., Kawaji, H., Kaneshiro, K., Harada, Y., Hayashi, S., Shimazaki, A., Morisaki, T., Mori, H., Oda, Y., Chen, S., Moriyama, T., Shimizu, S., & Nakamura, M. (2021). Validity of the prognostication tool PREDICT version 2.2 in Japanese breast cancer patients. Cancer Medicine, 10(5), 1605–1613. https://doi.org/10.1002/CAM4.3713

45. Zhong, X., Luo, T., Deng, L., Liu, P., Hu, K., Lu, D., Zheng, D., Luo, C., Xie, Y., Li, J., He, P., Pu, T., Ye, F., Bu, H., Fu, B., & Zheng, H. (2020). Multidimensional Machine Learning Personalized Prognostic Model in an Early Invasive Breast Cancer Population-Based Cohort in China: Algorithm Validation Study. JMIR Medical Informatics, 8(11), e19069. https://doi.org/10.2196/19069

46. Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing, 14(PART A), 99–108. https://doi.org/10.1016/J.ASOC.2013.07.016